yang_seonhyeHW28

Seonhye Yang

3/9/2019

Introduction

The dataset we are given looks at if set of people from the Pima Indian heritage have diabetes or not based on certain predictors. The Pima Indians live in Arizona, and are known to have a extremly high rate of diabetes. This is believed to be caused by a non-native diet and their water source and genetic predispostion. In fact, they have the highest prevalance of type 2 diabetes in the whole world.

Type 2 diabetes is due to insulin resistance, and causes the body to stop developing insulin. This is mostly found in adults and is usually associated with high body weight and low exercise.

The dataset we are given looks at the number of times a participant has been pregant, their glucose concentration, their diastolic blood pressure, how thick their tricep skin fold is, their insulin level, their BMI, and a value computed from a diabetes pedigree function. Based on the information about diabetes and the Pima Indians, it would be safe to guess that BMI, blood pressure, insulin level and age are probably good predictors for the glucose level of a Pima Indian.

The Data

Preparing the data

```
library(faraway)
data(pima)
pima <- pima[,-9]
pima = pima[which(pima$bmi!=0 &
                    pima$insulin!=0 &
                    pima$diastolic!=0 &
                    pima$glucose!=0 &
                    pima$triceps!=0),]
```

In this step, we have loaded our data and removed the "test" column, as we were specifically instructed not to use this. In addition to this, we have removed any rows that have zeroed out data points, as these will not help in the creation of our model.

Summaries of Data

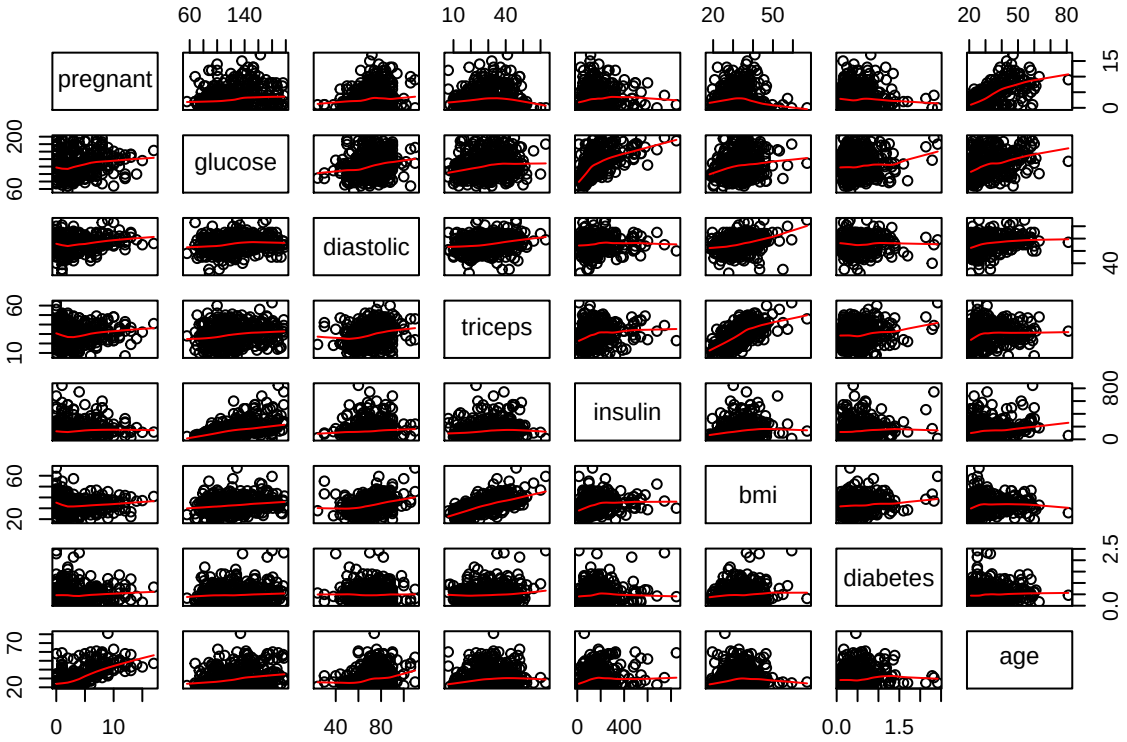
```
summary(pima)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
```

```
##     insulin           bmi             diabetes           age       
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00  
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00  
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00  
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86  
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00  
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

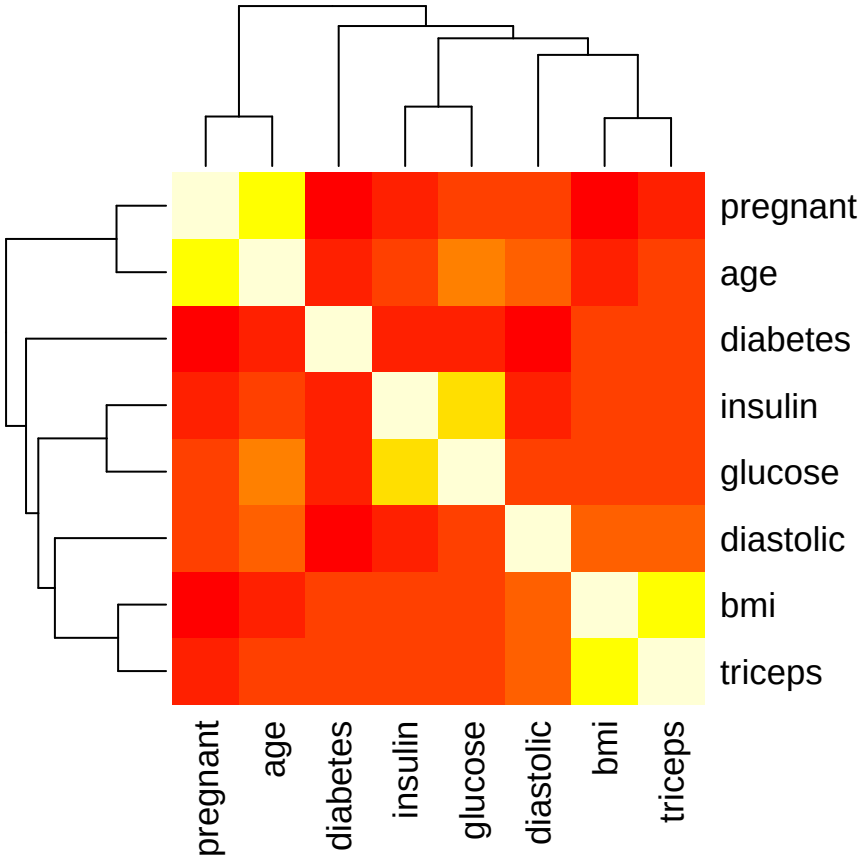
From this simple summary, we can try and see the grouping of the data within variables. "pregnant" is a quite widely varying column, and might need to be removed due to its range. However, the same can ve said about "age", yet we may wish to use this as a predictor, even though it has a huge range.

```
pairs(pima, panel = panel.smooth)
```



From this summary, we can try and glean what variables might be highly correlated. It would seem that insulin is quite correlated with glucose, so that is probably a good variable for us to begin with. In additio to his, triceps seems to be correlated with BMI, and age might have something to do with diastolic. It is important to know this as we would want to try and avoid interference between variables.

```
heatmap(cor(pima), symm=TRUE)
```



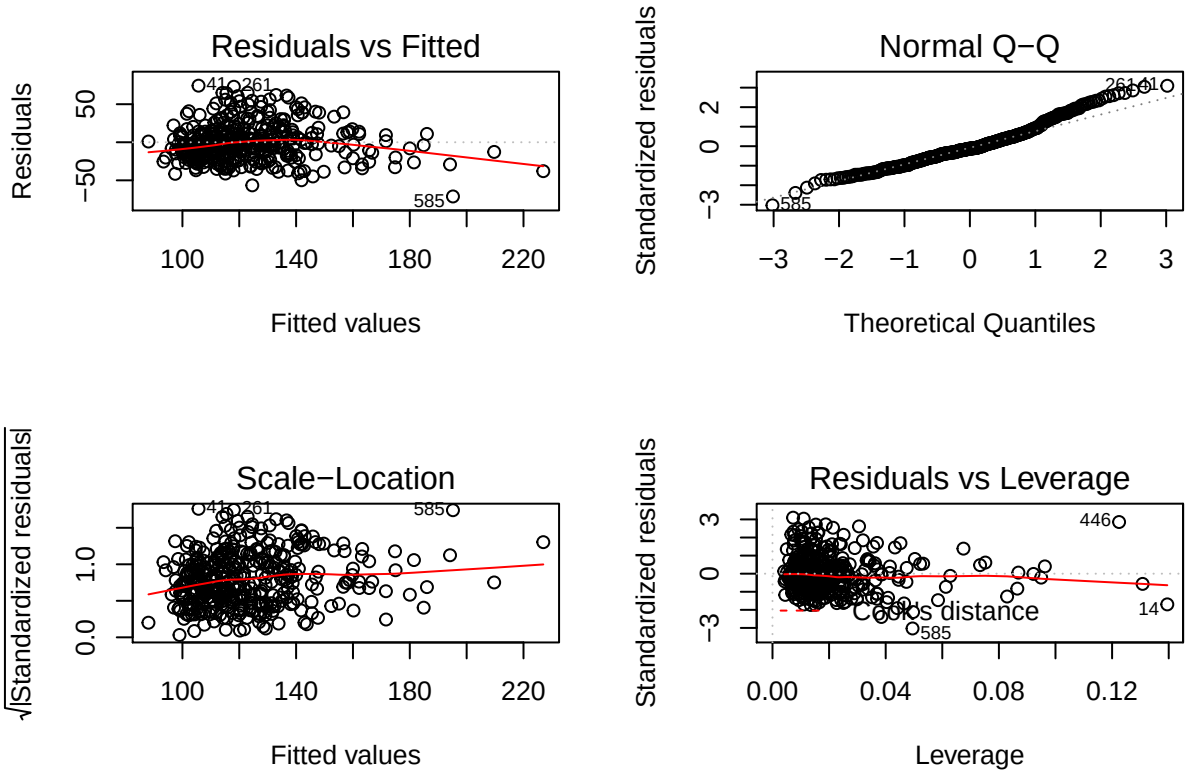
Looking at this heatmap of correlations, we can see that glucose and age may be more correlated than we previously throught. In addition, age and diastolic are quite correlated as well. All of these are important factors to keep in mind.

```
fit<- lm(glucose ~ ., data=pima)
summary(fit)
```

```
##
## Call:
## lm(formula = glucose ~ ., data = pima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.185 -15.558  -3.087  11.847  74.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.03002    8.44119   7.112 5.65e-12 ***
## pregnant     0.07383    0.52315   0.141 0.887848
## diastolic    0.21341    0.10769   1.982 0.048219 *
## triceps      0.07433    0.15769   0.471 0.637628
## insulin      0.13321    0.01084  12.293  < 2e-16 ***
## bmi          0.13038    0.24389   0.535 0.593239
## diabetes     4.17855    3.62412   1.153 0.249635
## age          0.57734    0.17182   3.360 0.000857 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.1 on 384 degrees of freedom
```

```
## Multiple R-squared:  0.4012, Adjusted R-squared:  0.3903
## F-statistic: 36.76 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(fit)
```



According to this summary, triceps, insulin, bmi, diabetes and age has p-value smaller than 0.05 and so they are significant. Besides that, both the adjusted r-square and the F test indicate the invalidity of the multiple linear regression model.

Searching for outliers

```
#testing for outliers
rfit<-rstudent(fit)
cValBonfer <- abs(qt(0.05/(2*nobs(fit)), df=df.residual(fit)-1, lower.tail=FALSE))

rfit[abs(rfit) > cValBonfer]
```

```
## named numeric(0)
```

The Bonferroni test shows that after our filtering, there are no outliers present in our data. This means that we don't have to remove data points to get a good fit.

```
#testing for influlential points
cdist <- cooks.distance(fit)
cdist[cdist >= 1]
```

```
## named numeric(0)
```

In addition to the Bonferroni test, we can try and find points that have a large cook distance based on our current simple model. As the results show, there are no points with high leverage, so we can leave our filtered

dataset intact for the time being.

Method

To generate a model that is better than the baseline model, we can begin by going directly to using stepwise variable selection and using the AIC metric to find a model with good fit:

```
steps <- step(fit)
```

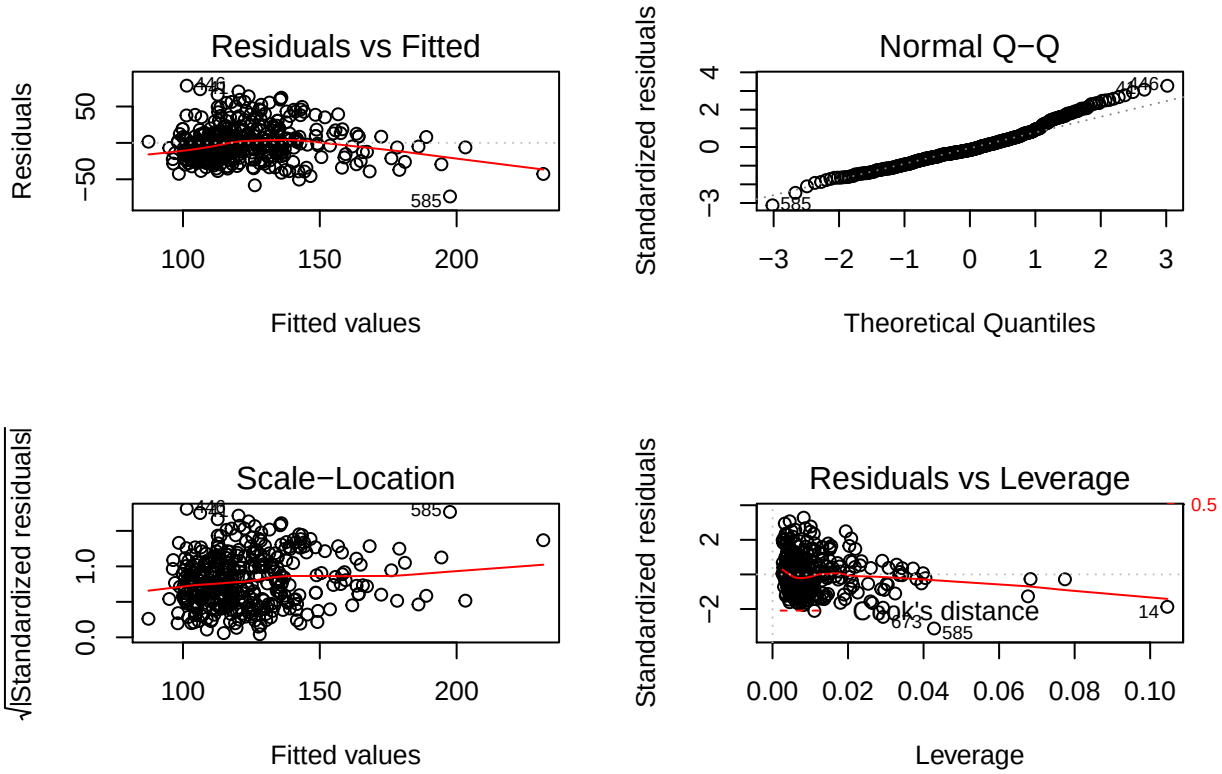
```
## Start:  AIC=2502.67
## glucose ~ pregnant + diastolic + triceps + insulin + bmi + diabetes +
##     age
##
##             Df Sum of Sq    RSS    AIC
## - pregnant   1        12 222986 2500.7
## - triceps    1       129 223104 2500.9
## - bmi        1       166 223141 2501.0
## - diabetes   1       772 223747 2502.0
## <none>                   222975 2502.7
## - diastolic  1      2280 225255 2504.7
## - age        1      6556 229530 2512.0
## - insulin    1     87755 310730 2630.8
##
## Step:  AIC=2500.69
## glucose ~ diastolic + triceps + insulin + bmi + diabetes + age
##
##             Df Sum of Sq    RSS    AIC
## - triceps    1       134 223120 2498.9
## - bmi        1       159 223145 2499.0
## - diabetes   1       765 223751 2500.0
## <none>                   222986 2500.7
## - diastolic  1      2297 225283 2502.7
## - age        1     12116 235102 2519.4
## - insulin    1     88060 311046 2629.2
##
## Step:  AIC=2498.93
## glucose ~ diastolic + insulin + bmi + diabetes + age
##
##             Df Sum of Sq    RSS    AIC
## - bmi        1       653 223773 2498.1
## - diabetes   1       808 223928 2498.3
## <none>                   223120 2498.9
## - diastolic  1      2297 225417 2500.9
## - age        1     12769 235889 2518.7
## - insulin    1     88093 311213 2627.4
##
## Step:  AIC=2498.07
## glucose ~ diastolic + insulin + diabetes + age
##
##             Df Sum of Sq    RSS    AIC
## - diabetes   1      1071 224845 2497.9
## <none>                   223773 2498.1
## - diastolic  1      3434 227207 2502.0
```

```
## - age        1     12396 236169 2517.2
## - insulin    1     94915 318688 2634.7
##
## Step:  AIC=2497.95
## glucose ~ diastolic + insulin + age
##
##             Df Sum of Sq    RSS    AIC
## <none>                   224845 2497.9
## - diastolic  1      3258 228103 2501.6
## - age        1     12964 237809 2517.9
## - insulin    1     98886 323731 2638.8
```

```
summary(steps)
```

```
##
## Call:
## lm(formula = glucose ~ diastolic + insulin + age, data = pima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.576 -15.015  -3.763  12.144  78.621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.4659     7.2187   9.069  < 2e-16 ***
## diastolic     0.2423     0.1022   2.371   0.0182 *
## insulin       0.1372     0.0105  13.063  < 2e-16 ***
## age           0.6036     0.1276   4.730 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.07 on 388 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3915
## F-statistic: 84.87 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(steps)
```



We can see from this simple AIC Step model, we have increased our R-squared to 0.3915, a meager improvement. We can run our outlier and leverage tests again to see if we should remove datapoints:

```
#testing for outliers
rfit<-rstudent(steps)
cValBonfer <- abs(qt(0.05/(2*nobs(steps)), df=df.residual(steps)-1, lower.tail=FALSE))

rfit[abs(rfit) > cValBonfer]
```

```
## named numeric(0)
```

```
#testing for influlential points
cdist <- cooks.distance(steps)
cdist[cdist >= 1]
```

```
## named numeric(0)
```

Again, we find no points with high leverage or being classified as outliers based on the Bonferroni test, so we can assume this is the best model we can make to predict glucose.

Results

To summarize our results and search, we came up with the following model:

$$Glucose(D, I, A) = \beta_1 D + \beta_2 I + \beta_3 A + \epsilon$$

Where D is the diastolic blood pressure, I is the insulin level, and A is the age. These variables were selected using a Stepwise AIC variable elimination to get rid of variables that were not statistically significant when trying to predict the glucose level of a Pima Indian. The following is a summary of our model:

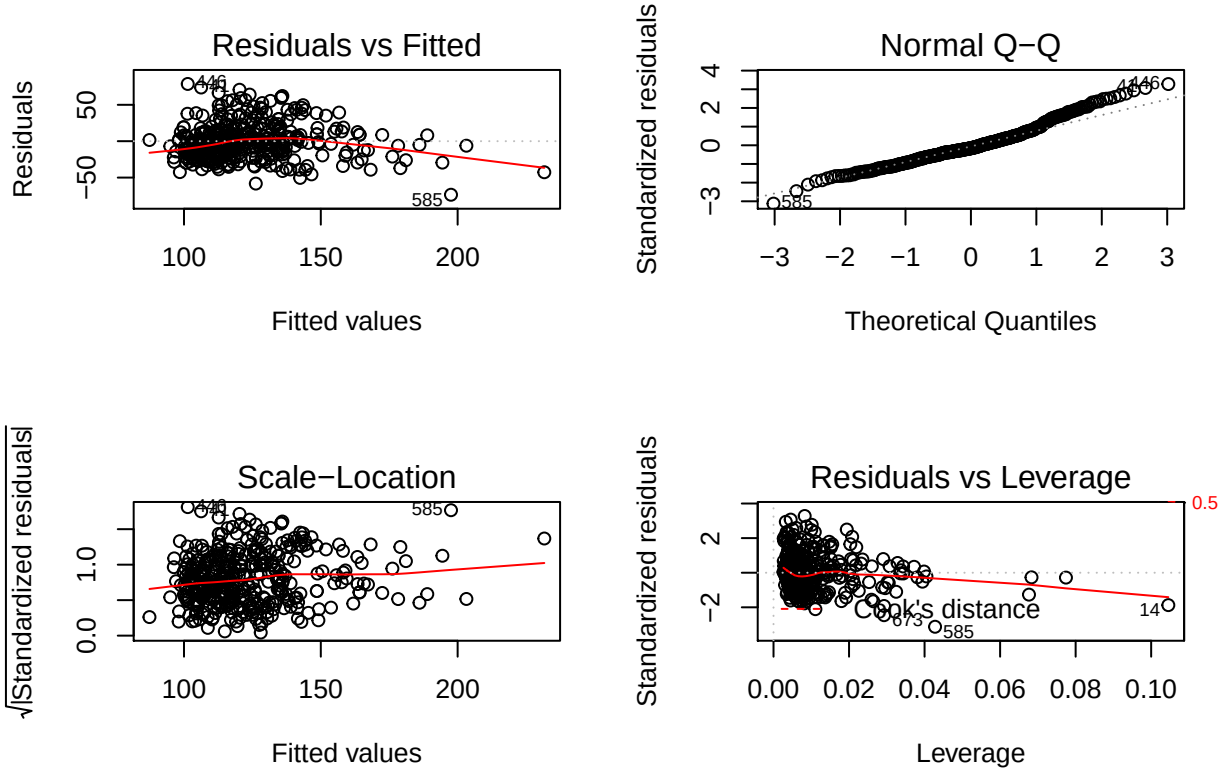
```
summary(steps)
```

```
##
## Call:
## lm(formula = glucose ~ diastolic + insulin + age, data = pima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.576 -15.015  -3.763  12.144  78.621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.4659     7.2187   9.069  < 2e-16 ***
## diastolic     0.2423     0.1022   2.371   0.0182 *
## insulin       0.1372     0.0105  13.063  < 2e-16 ***
## age           0.6036     0.1276   4.730 3.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.07 on 388 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3915
## F-statistic: 84.87 on 3 and 388 DF,  p-value: < 2.2e-16
```

This summary shows that all of our selected variables had a p values of 0.05 or less, making them significant. Something of interest is that both age and insulin have p values of less than 0.00001, making them super significant. As previosuly stated, this model has an R-Squared of 0.3915, indicating that it may not be the best fit for this data, and suggesting that a non-linear model may do better.

We can also look at diagnostic plots of our model:

```
par(mfrow = c(2,2))
plot(steps)
```



Looking at this set of diagnositc plots, we can see that this is a good model. Quickly summarizing these plots, nearly flat trendlines in both the residuals vs fitted and scale-location plus relatively evenly distributed points shows us that our residuals are evenly spaced from the line, indicating a linear dataset and a good fit. The linear Normal Q-Q plot also informs us of linearity and the Leverage plot indicates that point 14 may be out of line, but with a leverage of 0.10 this is hardly noticeable. So overall, our model is good.

Looking qualatatively at the predictors themselves, it is not surprising to see that both your blood pressure and the amount of insulin in your body help predict how much glucose is in one's system. Age was a slightly surprising factor, and looking at the coefficient of age, it is the most influencial of thse predictors. This indicates that as age increases, your ablity to produce glucose, or hold glucose increases as well.

Summary

All in all, we have created a model that can predict the amount of glucose in a Pima Indian using only insulin levels, diastolic blood pressure, and age. However, this model is not a excellent fit for the data, and leads us to believe that the data may require a more complex model to be fit to it.